CHAPTER 12
STANDARD SETTING

The Maine Department of Education, in an 18-month process with extensive input from educators and policy makers throughout the state, created four performance levels to describe student achievement:

- Does Not Meet the Standards,
- Partially Meets the Standards,
- Meets the Standards, and
- Exceeds the Standards.

Four policy considerations the department set for performance standards were that they be:

- concrete,
- consistent,
- challenging, and
- obtainable.

The process used to determine the MEA scores necessary for each performance level was developed with these policy considerations in mind. Two sources of data were gathered.

- Twenty-one panels consisting of about 300 educators, parents, businesspeople, and policy makers systematically looked at samples of student work and rated the work against the four Maine performance level descriptors.
- About 5,000 additional teachers rated student classroom work against those same performance level descriptors.

The results of these two approaches were averaged and then adjusted to minimize any inconsistency of the standards across the different grade levels. This last adjustment was accomplished by averaging the results for each grade with the results for the other two grades. The effect of this adjustment was kept small by counting the results of the grade under consideration four times as heavily as the results of either other grade.

PERFORMANCE LEVELS DEFINITIONS

The following charts contain the subject-specific performance level definition.

CHART 12-1
READING

Exceeds the Standards—The quality of a student's work at this level of proficiency exceeds the standards of performance as identified for Maine's *Learning Results* in English language arts (reading). The work demonstrates exemplary accomplishment in the comprehension of literary and informational texts, in the use of the skills and strategies of reading to answer questions, and in the demonstration of understanding of how words and images communicate. (Scaled scores: 561–580.)

Meets the Standards—The quality of a student's work at this level of proficiency meets the standards of performance as identified for Maine's *Learning Results* in English language arts (reading). The work demonstrates a consistent accomplishment in the comprehension of literary and informational texts, in the use of the skills and strategies of reading to answer questions, and in the demonstration of understanding of how words and images communicate. (Scaled scores: 541–560.)

Partially Meets the Standards—The quality of a student's work at this level of proficiency partially meets the standards of performance as identified for Maine's *Learning Results* in English language arts (reading). The work demonstrates inconsistent accomplishment in the comprehension of literary and informational texts, in the use of the skills and strategies of reading to answer questions, and in the demonstration of understanding of how words and images communicate. (Scaled scores: 521–540.)

Does Not Meet the Standards—The quality of a student's work at this level of proficiency does not meet the standards of performance as identified for Maine's *Learning Results* in English language arts (reading). The work demonstrates limited accomplishment in the comprehension of literary and informational texts, in the use of the skills and strategies of reading to answer questions, and in the demonstration of understanding of how words and images communicate. (Scaled scores: 501-520.)

CHART 12-2
WRITING

Exceeds the Standards—The quality of a student's written compositions at this level of proficiency exceeds the standards of performance as identified for Maine's *Learning Results* in English language arts (writing). The student's work demonstrates exemplary accomplishment in both the development of the topic/idea and the use of Standard English conventions in first-draft writing. (Scaled scores:561–580.)

Meets the Standards—The quality of a student's written compositions at this level of proficiency meets the standards of performance as identified for Maine's *Learning Results* in English language arts (writing). The student's work demonstrates proficiency in both the development of the topic/idea and the use of Standard English conventions in first-draft writing. (Scaled scores:541–560.)

Partially Meets the Standards—The quality of a student's written compositions at this level of proficiency partially meets the standards of performance as identified for Maine's *Learning Results* in English language arts (writing). The student's work demonstrates writing skills that may show moderate development of topic/ideas and/or some errors in Standard English conventions that may interfere with communication. (Scaled scores:521–540.)

Does Not Meet the Standards—The quality of a student's written compositions at this level does not meet the standards of performance as identified for Maine's *Learning Results* in English language arts (writing). The student's work demonstrates writing skills that show limited development of topic/idea and/or many errors in Standard English conventions that interfere with communication of ideas. (Scaled scores:501–520.)

CHART 12-3
HEALTH EDUCATION

Exceeds the Standards—The quality of a student's work at this level of proficiency exceeds the standards of performance as identified for Maine's *Learning Results* in health education. The student demonstrates exemplary knowledge of content and skills related to health promotion and disease prevention including communication, decision making, analysis, and risk reduction. (Scaled scores:561–580.)

Meets the Standards—The quality of a student's work at this level of proficiency meets the standards of performance as identified for Maine's *Learning Results* in health education. The student demonstrates consistent knowledge of content and skills related to health promotion and disease prevention including communication, decision making, analysis, and risk reduction. (Scaled scores:541–560.)

Partially Meets the Standards—The quality of a student's work at this level of proficiency partially meets the standards of performance as identified for Maine's *Learning Results* in health education. The student demonstrates partial and/or inconsistent knowledge of content and skills related to health promotion and disease prevention including communication, decision making, analysis, and risk reduction. (Scaled scores:521–540.)

Does Not Meet the Standards—The quality of a student's work at this level of proficiency does not meet the standards of performance as identified for Maine's *Learning Results* in health education. The student demonstrates a limited knowledge of content and skills related to health promotion and disease prevention including communication, decision making, analysis, and risk reduction. (Scaled scores:501–520.)

CHART 12-4
MATHEMATICS

Exceeds the Standards—The quality of a student's work at this level of proficiency exceeds the standards of performance as identified for Maine's *Learning Results* in mathematics. The student's overall performance demonstrates exemplary knowledge of content, process, problem-solving, and communication skills. (Scaled scores:561–580.)

Meets the Standards—The quality of a student's work at this level of proficiency meets the standards of performance as identified for Maine's *Learning Results* in mathematics. The student's work consistently shows complete knowledge of mathematical content, process, reasoning, and communication skills, as well as problem-solving abilities. (Scaled scores:541–560.)

Partially Meets the Standards—The quality of a student's work at this level of proficiency partially meets the standards of performance as identified for Maine's *Learning Results* in mathematics. The student's work demonstrates a partial and/or inconsistent knowledge of mathematical content, process, reasoning, and communication skills, and problem-solving abilities. (Scaled scores:521–540.)

Does Not Meet the Standards—The quality of a student's work at this level of proficiency does not meet the standards of performance as identified for Maine's *Learning Results* in mathematics. The student's work demonstrates a limited knowledge of mathematical content, process, reasoning, and communication skills, as well as problem-solving ability. (Scaled scores:501–520.)

CHART 12-5
SCIENCE & TECHNOLOGY

Exceeds the Standards—The quality of a student's work at this level of proficiency exceeds the standards of performance as identified for Maine's *Learning Results* in science and technology. The student demonstrates exemplary knowledge of content including life, physical, and earth/space sciences and scientific inquiry, reasoning, and communication skills. (Scaled scores:561–580.)

Meets the Standards—The quality of a student's work at this level of proficiency meets the standards of performance as identified for Maine's *Learning Results* in science and technology. The student demonstrates consistent knowledge of content including life, physical, and earth/space sciences and scientific inquiry, reasoning, and communication skills. (Scaled scores:541–560.)

Partially Meets the Standards—The quality of a student's work at this level of proficiency partially meets the standards of performance as identified for Maine's *Learning Results* in science and technology. The student demonstrates partial and/or inconsistent knowledge of content including life, physical, and earth/space sciences and scientific inquiry, reasoning, and communication skills. (Scaled scores:521–540.)

Does Not Meet the Standards—The quality of a student's work at this level of proficiency does not meet the standards of performance as identified for Maine's *Learning Results* in science and technology. The student demonstrates limited knowledge of content including life, physical, and earth/space sciences and scientific inquiry, reasoning, and communication skills. (Scaled scores:501–520.)

CHART 12-6
SOCIAL STUDIES

Exceeds the Standards—The quality of a student's work at this level of proficiency exceeds the standards of performance as identified for Maine's *Learning Results* in social studies. The student demonstrates exemplary knowledge of content of major social studies concepts, consistently applies complex thinking skills, and communicates ideas clearly in all situations. (Scaled scores:561–580.)

Meets the Standards—The quality of a student's work at this level of proficiency meets the standards of performance as identified for Maine's *Learning Results* in social studies. The student demonstrates consistent knowledge of content of major social studies concepts, usually applies complex thinking skills, and communicates ideas clearly in most situations. (Scaled scores:541–560.)

Partially Meets the Standards—The quality of a student's work at this level of proficiency partially meets the standards of performance as identified for Maine's *Learning Results* in social studies. The student demonstrates some knowledge of content of major social studies concepts, inconsistently applies complex thinking skills, and communicates ideas clearly in some situations. (Scaled scores:521–540.)

Does Not Meet the Standards—The quality of a student's work at this level of proficiency does not meet the standards of performance as identified for Maine's *Learning Results* in social studies. The student demonstrates a limited knowledge of content of major social studies concepts, does not apply complex thinking skills, and communicates ideas clearly in few or no situations. (Scaled scores:501–520.)

CHART 12-7
VISUAL & PERFORMING ARTS

Exceeds the Standards—The quality of a student's work at this level of proficiency exceeds the standards of performance as identified for Maine's *Learning Results* in visual and performing arts. The student demonstrates exemplary knowledge of content and application of skills of the visual and performing arts, including creative expression, cultural heritage, and criticism and aesthetics. (Scaled scores:561–580.)

Meets the Standards—The quality of a student's work at this level of proficiency meets the standards of performance as identified for Maine's *Learning Results* in visual and performing arts. The student demonstrates consistent knowledge of content and application of skills of the visual and performing arts, including creative expression, cultural heritage, and criticism and aesthetics. (Scaled scores:541–560.)

Partially Meets the Standards—The quality of a student's work at this level of proficiency partially meets the standards of performance as identified for Maine's *Learning Results* in visual and performing arts. The student demonstrates partial and/or inconsistent knowledge of content and application of skills of the visual and performing arts, including creative expression, cultural heritage, and criticism and aesthetics. (Scaled scores:521–540.)

Does Not Meet the Standards—The quality of a student's work at this level of proficiency does not meet the standards of performance as identified for Maine's *Learning Results* in visual and performing arts. The student demonstrates limited knowledge of content and application of skills of the visual and performing arts, including creative expression, cultural heritage, and criticism and aesthetics. (Scaled scores:501–520.)

STANDARD SETTING METHODS

There were two standard setting methods used for the MEA: the Body of Work (BoW) method (Kingston, Kahl, Sweeney, & Bay, 2000) and the Contrasting Group (CG) method (Livingston & Zieky, 1982). Threshold scores resulting from the two methods were aggregated to obtain the minimum scores for each performance level.

The two methods and their implementations are described in the following sections. The threshold scores that were recommended to and accepted by the DOE are also presented.

CONTRASTING GROUP (CG)

The contrasting group method is based on the notion that examinees can be divided into two contrasting groups (Livingston & Zieky, 1982). For example, for the MEA, these two groups could be the group of examinees that Meets the Standards (this includes those who Exceeds the Standards) and the group of students that do not (this includes those who Partially Meets the Standards and those in the Does not Meet the Standards categories).

Prior to the implementation of the BoW standard setting method, student rosters were sent to select schools with a request for teachers to assign performance levels to selected students in different subject areas. The instructions given to the teachers were as follows:

1. Carefully review the Maine *Learning Results* for this content area.
2. Carefully review the performance level definitions.
3. For each student listed, indicate the performance level that matches the student's achievement of the Maine *Learning Results*. (1 = Exceeds the Standards; 2 = Meets the Standard; 3 = Partially Meets the Standard; 4 = Does Not Meet the Standard)
4. Return the completed form to your building principal.

Included in the instructions is the information that the task of assigning performance levels was to be performed by the teacher who is currently teaching or who most recently taught this content area to the identified student. Teachers and principals involved in this study were told that information collected will be used along with information collected during standard setting sessions on July 26-29, 1999, to establish the performance level cutscores for the MEA.

A total of 73 schools in Maine were selected and asked to participate in this study: 44 for grade 4, 12 for grade 8, and 17 for grade 11, across the six subject areas. The number of students selected for this study for each grade level and subject combination is presented in Table 12-1. These are the numbers of students that teachers have to assign to different performance levels.

Data collected from this effort were analyzed to obtain threshold scores for each performance level in each grade and subject area. These thresholds were combined with thresholds resulting from the BoW method to

obtain the final thresholds recommended to the DOE. The method of combining the thresholds is discussed later in this chapter.

Table 12-1 Number of Selected Students for the Contrasting Group			
Subject	**Grade 4**	**Grade 8**	**Grade 11**
Reading	330	340	328
Mathematics	328	326	338
Science and Technology	314	333	330
Social Studies	315	330	330
Health Education	312	332	357
Visual and Performing Arts	310	379	381

BODY OF WORK (BOW)

On July 26-29, 1999, panels were assembled for the implementation of the Body of Work (BoW) standard-setting method. The hallmark of the BoW method is that panelists examine complete student response sets (student responses to multiple-choice questions and samples of actual student work on open-response questions) and match each student response set to one of the MEA performance level categories. This is done in three major steps: (1) training/calibration, (2) range finding, and (3) pinpointing.

TRAINING/CALIBRATION

During this first phase of the MEA standard-setting process, panelists reviewed all MEA test questions for their assigned content area and grade level, and content- and grade-specific descriptors for each performance level. Panelists were given the opportunity to discuss and comment on test questions and descriptors. Next, to ensure that panelists attained a common interpretation of performance descriptors and the relationship of those descriptors to student work, panel members individually assigned performance levels to a set of six sample student responses. Panelists then compared their individual results and discussed at length how the performance level descriptors supported their conclusions.

RANGE-FINDING

During the range-finding phase of standard setting, identical sets of student work that spanned the score continuum were provided to each panelist. Panelists were asked to independently categorize the sets as Exceeds the Standards, Meets the Standards, Partially Meets the Standards, or Does Not Meet the Standards, based on the performance level descriptors. This process revealed which levels of student work generated the most agreement and which generated the most disagreement among panelists. The results were documented, and the levels of the sets of work that generated the most disagreement defined the score intervals in which the threshold scores must fall.

PINPOINTING

Additional sets of student work from score ranges that generated disagreement were presented to panelists. Panelists assigned performance levels to these sets of responses. The minimum score for each performance level was precisely pinpointed by determining the score around which there was, collectively, the maximum disagreement between panelists. This is the point that best represents the transition from response sets at a higher level to those at a lower level.

PANELISTS

Twenty-one panels were convened to set performance standards for the MEA—one panel for each grade level (4, 8, and 11) in seven areas—(1) reading, (2) writing, (3) mathematics, (4) science, (5) social studies, (6) health, and (7) visual and performing arts. The panels were composed of educators, parents and business leaders, and members of the general public.

IMPLEMENTATION

Following is a detailed description of the steps followed in implementing the MEA BoW standard-setting design.

BEFORE THE MEETING

1. For each subject-grade combination (e.g., grade 8 mathematics) pinpointing folders were prepared from samples of student work. This sample was double-scored to increase the accuracy of the standard-setting process. Any students whose body of work was of uneven quality (for example, some open-response questions with scores of four and others with scores of one) were excluded, as were students whose open-response and multiple-choice responses were particularly discrepant. Folders ranged in scores from the highest obtained score in the

remaining sample to the "approximately chance level" (0.25 times the number of multiple-choice items plus one times the number of open-response items). Each folder consisted of five sets of student work at each of four score points (e.g., five 12s, five 13s, five 14s, and five 15s), with the exception of the top folder (folder with highest scores). The top folder differed because there often were fewer than five papers available at any particular score point. Thus, the twenty papers in the top folder covered a wider range of scores. Approximately ten pinpointing folders were created for each subject-grade combination.

2. Range-finding folders were prepared from the pinpointing folders. The highest-scoring and two lowest-scoring papers were selected from each pinpointing folder. Thus, range-finding folders had about thirty samples of student work in each.
3. For each subject-grade combination, six student response sets spanning the range of performance were identified from the pinpointing folders. The facilitator reviewed the sets and prepared training notes consisting of points to be made during discussion of those student response sets. Focus was on ways responses illustrate characteristics described in the performance level definitions.
4. The Maine Department of Education created a list of members of each panel (one panel per subject area, four subject areas per grade, and three grades), ensuring each group had the proper diversity of membership (educator, parent, policy-maker, businessperson, ethnicity, gender, etc.). Color-coded name tags were provided to panel members.

GENERAL MEETING

Before the panels broke into separate groups, there was a general session at which logistical issues were addressed and the standard-setting procedures explained by the chief of standard setting. Major steps of the panel meeting portion of the meeting were described.

PANEL MEETING

1. Facilitators distributed the descriptor of a four-point response to each open-response question. Panel members were asked to review and discuss the test questions—open-response and multiple-choice.

(Panelists had been asked to answer the questions before the meeting, and they were to have brought with them the tests and the performance level definitions. Additional copies were distributed to those who needed them.)

2. The facilitators led a discussion of the performance level definitions.
3. Training folders were distributed to every judge. The multiple-choice display at the end of a set was pointed out. Facilitators explained that it too should be considered when judgments are being made about the student work.
4. Judges were asked to rank independently the six previously identified student response sets based on overall quality, keeping in mind the performance level descriptions. Each judge listed the six student serial numbers in rank order from high to low performance on a separate piece of paper.
5. While the judges rank ordered the six student response sets, the facilitator wrote the serial numbers of the six sets on an overhead transparency in a vertical list in order from highest performance to lowest performance. When the judges completed their rankings, the facilitators showed the score rankings on the overhead projector and had the judges note the extent of agreement.
6. Judges were asked to assign each of the six response sets to a performance level. They each wrote the performance level initials (E, M, P, or D) next to the student serial numbers they listed in rank order in step 4.
7. Facilitators drew four columns to the right of the six serial numbers on the overhead transparency, and labeled the columns E, M, P, and D. Facilitators recorded the judges' ratings (based on shows of hands) next to the serial numbers on the overhead.
8. Facilitators led a discussion of the six response sets as they related to the performance levels.

9. The heterogeneous (range-finding) folders were distributed to every judge. The facilitators pointed out the multiple-choice display at the end of a set, and explained that it too should be considered when judgments are being made about the student work.
10. Facilitators distributed a Range-Finding Rating Form to each judge, and asked the judges to enter their names in the name boxes and encode a home telephone number in the "ID" field. Judges were given the opportunity to reconsider their ratings of the six student response sets and transfer their "final" ratings to the Range-Finding Rating Form on which the serial numbers for these and other response sets in the heterogeneous folder had been entered in order from high to low performance.
11. Judges were asked to decide independently the performance levels of the rest of the sets in the heterogeneous folder and record their ratings on their Range-Finding Rating Forms in the left set of columns.
12. Judges' ratings were recorded on the "Range-Finding" overhead transparency, based on shows of hands. Judges were asked to view the overhead and decide if they wanted to change their minds regarding any of the student response sets. Group discussion was allowed. Changed ratings were recorded in the "Second Ratings" columns of the Range-Finding Rating Form.
13. When the judges completed step 12, their materials were collected. From these data, the chief of standard setting determined the pinpointing folder or folders that must be evaluated by the judges for determining each of the three cut points.
14. For each pinpointing folder, the decision to be made for each folder was indicated, e.g.,

    Folders 3 and 4—E or M?

    Folders 9 and 10—M or P?

    Folder 15—P or D?

15. The group of judges was divided into thirds. Each small group examined the folder or folders for one cut score[1]. Each judge independently completed a Pinpointing Rating Form, including the name boxes and ID field, for each folder he or she was assigned. Materials were rotated so all three small groups examined the folder or folders for every cut point.
16. All standard-setting materials (ranking sheets, forms, folders, tests, definitions, etc.) were collected and returned to the chief of standard setting.

As panelists turned in their materials, they were given an evaluation form to fill out and were invited to return later to see a summary of the results.

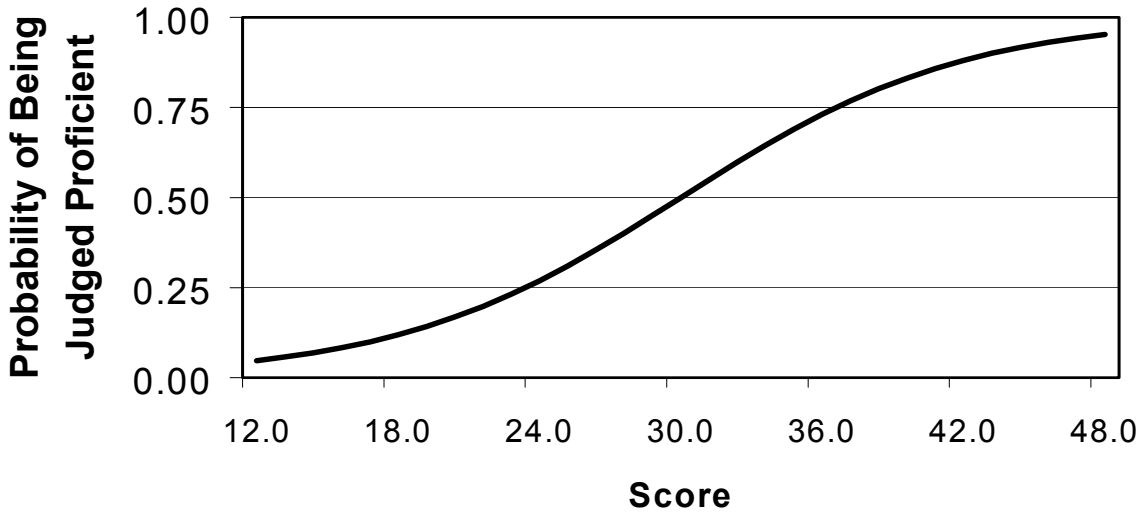
[1] The purpose of dividing the group into thirds was to reduce the need for multiple copies of folders. This way, each group worked with one-third of the folders, finished the work on one cut score, and then passed the folders to the next group for them to do the same.

DATA ANALYSIS

Data collected from CG and BoW were analyzed separately using logistic regression. Using data collected through each method, a separate logistic regression was run for each threshold decision. The unit of analysis for the CG data was a teacher's decision regarding each student. For the BoW data, the unit of analysis is a panelist's decision about a single student's body of work. Test scores were used to predict the probability of a student's work being classified as meeting or exceeding each performance level. Figure 12-1 provides a graphical example of the results of a logistic regression.

Figure 12-1

Graphical Example of Logistic Regression Results



Note, in Figure 12-1, it is at a test score of thirty that the probability of being judged Meets the Standards is 0.5. Thus, thirty would be the minimum score at which a student would be considered Meets the Standards.

A separate regression analysis was done for each performance level for each grade and subject combination based on each set of collected data from CG and BoW methods. Each threshold score computed was associated with a standard error. Standard errors were estimated by applying the logistic regression technique separately to each panelist's or teacher's data. Thus, for each threshold decision, there was a distribution of estimated

thresholds. The standard error was estimated as the standard deviation of the distribution divided by the square root of the number of panelists (for BoW) or teachers (for CG).

RESULTS

Threshold scores resulting from each method were presented to the DOE along with their associated standard errors as described above. A decision was made to combine the corresponding thresholds and smooth them across grades. The following steps outline the manner by which the final cutpoints were computed.

1. Based on the actual distribution of scores of students who took the tests, each cutpoint was converted to a z-equivalent score.
2. The z-equivalent scores of the BoW and CG cutpoints were combined by computing the weighted average (BoW:CG::2:1). This was done for each pair of performance level threshold for each subject area for each grade.
3. The corresponding z-equivalent cutpoints for each subject area for each performance level were "smoothed" across grades. This was done by computing the 4:1:1 weighted average of grade level cutpoints, where the cutpoint for the grade of interest is weighted four times as much as the cutpoints for the other two grades.
4. The resulting cutpoints (which are in z-equivalents score metric) are then converted to the raw score metric.

Table 12-2 presents the final threshold determinations that were used to report results from the 1999 administration of the MEA.

Table 12-2
Threshold (Minimum) Total Test Score For Each Performance Category

Grade	Subject Area	Maximum Score on Test	Threshold Score		
			Exceeds the Standards	Meets the Standards	Partially Meets the Standards
4	Reading	53	46.60	33.72	21.30
	Writing	30	26.64	18.56	9.91
	Mathematics	41	36.19	26.07	15.73
	Science	41	33.69	27.33	13.75
	Social Studies	39	32.16	25.31	17.44
	Health*	28	16.67	13.27	7.82
	Visual and Performing Arts*	28	13.75	10.35	6.81
8	Reading	52	44.91	33.10	21.14
	Writing	30	27.21	18.09	10.91
	Mathematics	41	37.30	24.40	12.23
	Science	41	33.71	25.99	16.03
	Social Studies	41	31.66	23.63	14.38
	Health*	28	20.37	13.15	5.68
	Visual and Performing Arts*	28	18.46	11.24	6.75
11	Reading	53	47.93	37.09	23.38
	Writing	30	26.96	20.12	12.09
	Mathematics	41	36.01	24.37	12.83
	Science	41	34.27	26.22	13.48
	Social Studies	39	30.66	21.00	12.76
	Health*	28	19.58	13.75	4.77
	Visual and Performing Arts*	28	20.18	14.59	9.50

*Information presented is based on the particular test forms used in standard setting.